Gaussian Processes: Basic Properties and GP Regression

Steffen Grünewälder

University College London

20. Januar 2010

Outline

Gaussian Random Variables

UCL

Definition

A Gaussian random variable X is completely specified by its mean μ and standard deviation σ. Its density function is:

$$\mathbf{P}[X = x] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Definition

A multivariate Gaussian random variable $\boldsymbol{X}$ is completely specified by its mean $\boldsymbol{\mu}$ and covariance matrix Σ (positive definite and symmetric). Its density function is:

$$\mathbf{P}[\boldsymbol{X} = \boldsymbol{x}] = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Gaussian Random Variables

UCL

Definition

A Gaussian random variable X is completely specified by its mean μ and standard deviation σ. Its density function is:

$$\mathbf{P}[X = x] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Definition

A multivariate Gaussian random variable $\boldsymbol{X}$ is completely specified by its mean $\boldsymbol{\mu}$ and covariance matrix Σ (positive definite and symmetric). Its density function is:

$$\mathbf{P}[\boldsymbol{X} = \boldsymbol{x}] = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Gaussian Process - Definition

Definition

A Gaussian process $f(x)$ is a collection of random variables, any finite number of which have a joint Gaussian distribution. A Gaussian process is completely specified by its mean function $\mu(x)$ and its covariance function $k(x, y)$. For $n \in \mathbb{N}$ and $x_1, \ldots, x_n$:

$$(f(x_1), \ldots, f(x_n)) \sim \mathcal{N}((\mu(x_1), \ldots, \mu(x_n)), \boldsymbol{K})$$

$$\boldsymbol{K} := \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & \ldots \\ k(x_2, x_1) & k(x_2, x_2) & \ldots \\ \ldots & & \end{pmatrix}$$

Sampling from a GP

- Goal: Generate a draw from a GP with mean μ and covariance $\boldsymbol{K}$.
- Compute Cholesky decomposition of $\boldsymbol{K}$, i.e.

$$\boldsymbol{K} = \boldsymbol{L}\boldsymbol{L}^\top,$$

and $\boldsymbol{L}$ is lower triangular.
- Generate

$$\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}).$$

- Compute

$$\boldsymbol{x} = \mu + \boldsymbol{L}\boldsymbol{u}.$$

- $\boldsymbol{x}$ has the right distribution, i.e.

$$\mathbf{E}(\boldsymbol{x} - \mu)(\boldsymbol{x} - \mu)^\top = \boldsymbol{L}\mathbf{E}[\boldsymbol{u}\boldsymbol{u}^\top]\boldsymbol{L}^\top = \boldsymbol{K}.$$

- Often numerical unstable: Add $\epsilon\boldsymbol{I}$ to the covariance.

Sampling from a GP

- Goal: Generate a draw from a GP with mean μ and covariance $\boldsymbol{K}$.
- Compute Cholesky decomposition of $\boldsymbol{K}$, i.e.

$$\boldsymbol{K} = \boldsymbol{L}\boldsymbol{L}^\top,$$

  and $\boldsymbol{L}$ is lower triangular.
- Generate

$$\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}).$$

- Compute

$$\boldsymbol{x} = \mu + \boldsymbol{L}\boldsymbol{u}.$$

- $\boldsymbol{x}$ has the right distribution, i.e.

$$\mathbf{E}(\boldsymbol{x} - \mu)(\boldsymbol{x} - \mu)^\top = \boldsymbol{L}\mathbf{E}[\boldsymbol{u}\boldsymbol{u}^\top]\boldsymbol{L}^\top = \boldsymbol{K}.$$

- Often numerical unstable: Add $\epsilon \boldsymbol{I}$ to the covariance.

Sampling from a GP

UCL

- Goal: Generate a draw from a GP with mean μ and covariance $\boldsymbol{K}$.
- Compute Cholesky decomposition of $\boldsymbol{K}$, i.e.

$$\boldsymbol{K} = \boldsymbol{L}\boldsymbol{L}^\top,$$

  and $\boldsymbol{L}$ is lower triangular.
- Generate

$$\boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}).$$

- Compute

$$\boldsymbol{x} = \mu + \boldsymbol{L}\boldsymbol{u}.$$

- $\boldsymbol{x}$ has the right distribution, i.e.

$$\mathbf{E}(\boldsymbol{x} - \mu)(\boldsymbol{x} - \mu)^\top = \boldsymbol{L}\mathbf{E}[\boldsymbol{u}\boldsymbol{u}^\top]\boldsymbol{L}^\top = \boldsymbol{K}.$$

- Often numerical unstable: Add $\epsilon\boldsymbol{I}$ to the covariance.

Sampling from a GP

- Goal: Generate a draw from a GP with mean μ and covariance $\boldsymbol{K}$.
- Compute Cholesky decomposition of $\boldsymbol{K}$, i.e.

$$\boldsymbol{K} = \boldsymbol{L}\boldsymbol{L}^\top,$$

  and $\boldsymbol{L}$ is lower triangular.
- Generate

$$\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}).$$

- Compute

$$\boldsymbol{x} = \mu + \boldsymbol{L}\boldsymbol{u}.$$

- $\boldsymbol{x}$ has the right distribution, i.e.

$$\mathbf{E}(\boldsymbol{x} - \mu)(\boldsymbol{x} - \mu)^\top = \boldsymbol{L}\mathbf{E}[\boldsymbol{u}\boldsymbol{u}^\top]\boldsymbol{L}^\top = \boldsymbol{K}.$$

- Often numerical unstable: Add $\epsilon \boldsymbol{I}$ to the covariance.

Sampling from a GP

- Goal: Generate a draw from a GP with mean μ and covariance $\boldsymbol{K}$.
- Compute Cholesky decomposition of $\boldsymbol{K}$, i.e.

$$\boldsymbol{K} = \boldsymbol{L}\boldsymbol{L}^\top,$$

  and $\boldsymbol{L}$ is lower triangular.
- Generate

$$\boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}).$$

- Compute

$$\boldsymbol{x} = \mu + \boldsymbol{L}\boldsymbol{u}.$$

- $\boldsymbol{x}$ has the right distribution, i.e.

$$\mathbf{E}(\boldsymbol{x} - \mu)(\boldsymbol{x} - \mu)^\top = \boldsymbol{L}\mathbf{E}[\boldsymbol{u}\boldsymbol{u}^\top]\boldsymbol{L}^\top = \boldsymbol{K}.$$

- Often numerical unstable: Add $\epsilon \boldsymbol{I}$ to the covariance.

Sampling from a GP

- Goal: Generate a draw from a GP with mean μ and covariance $\boldsymbol{K}$.
- Compute Cholesky decomposition of $\boldsymbol{K}$, i.e.

$$\boldsymbol{K} = \boldsymbol{L}\boldsymbol{L}^\top,$$

  and $\boldsymbol{L}$ is lower triangular.
- Generate

$$\boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}).$$

- Compute

$$\boldsymbol{x} = \mu + \boldsymbol{L}\boldsymbol{u}.$$

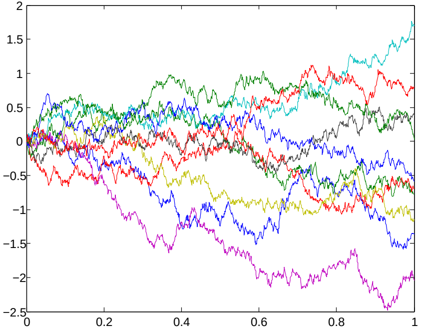
- $\boldsymbol{x}$ has the right distribution, i.e.

$$\mathbf{E}(\boldsymbol{x} - \mu)(\boldsymbol{x} - \mu)^\top = \boldsymbol{L}\mathbf{E}[\boldsymbol{u}\boldsymbol{u}^\top]\boldsymbol{L}^\top = \boldsymbol{K}.$$

- Often numerical unstable: Add $\epsilon \boldsymbol{I}$ to the covariance.

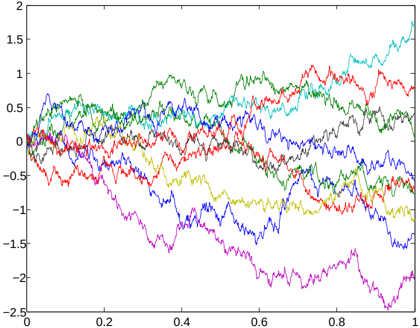
Example 1: Brownian Motion

- Most famous GP: Brownian Motion.
- Process on the real line starting at time 0 with value $f(0) = 0$.
- Covariance: $k(s, t) = \min\{s, t\}$.
- Brownian Motion is a Markov process. Means intuitively that for times $t_1 < t_2 < t_3$ the value of $f(t_3)$ conditional on $f(t_2)$ is independent of $f(t_1)$.



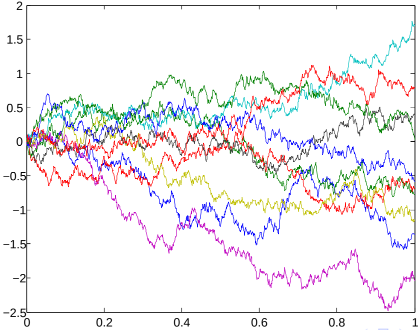
Example 1: Brownian Motion

- Most famous GP: Brownian Motion.
- Process on the real line starting at time 0 with value $f(0) = 0$.
- Covariance: $k(s, t) = \min\{s, t\}$.
- Brownian Motion is a Markov process. Means intuitively that for times $t_1 < t_2 < t_3$ the value of $f(t_3)$ conditional on $f(t_2)$ is independent of $f(t_1)$.



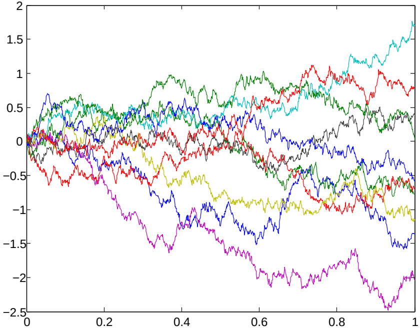
Example 1: Brownian Motion

- Most famous GP: Brownian Motion.
- Process on the real line starting at time 0 with value $f(0) = 0$.
- Covariance: $k(s, t) = \min\{s, t\}$.
- Brownian Motion is a Markov process. Means intuitively that for times $t_1 < t_2 < t_3$ the value of $f(t_3)$ conditional on $f(t_2)$ is independent of $f(t_1)$.



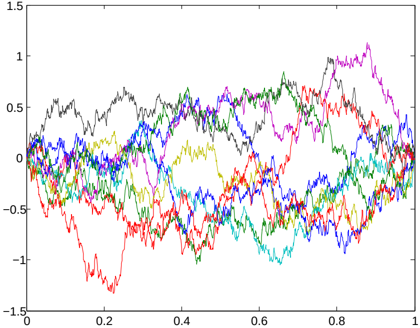
Example 1: Brownian Motion

- Most famous GP: Brownian Motion.
- Process on the real line starting at time 0 with value $f(0) = 0$.
- Covariance: $k(s, t) = \min\{s, t\}$.
- Brownian Motion is a Markov process. Means intuitively that for times $t_1 < t_2 < t_3$ the value of $f(t_3)$ conditional on $f(t_2)$ is independent of $f(t_1)$.



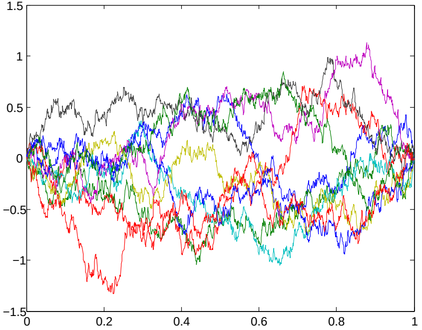
Example 2: Brownian Bridge

- A bridge is a stochastic process that is "clamped" at two points, i.e. each path goes (w.p. 1) through two specified points.
- Example: Brownian Bridge on $[0, 1]$ with $f(0) = f(1) = 0$.
- Covariance: $k(s, t) = \min\{s, t\} - st$



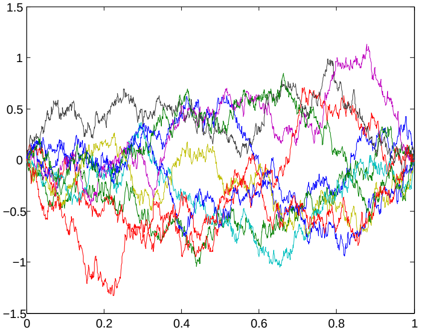
Example 2: Brownian Bridge

- A bridge is a stochastic process that is “clamped” at two points, i.e. each path goes (w.p. 1) through two specified points.
- Example: Brownian Bridge on $[0, 1]$ with $f(0) = f(1) = 0$.
- Covariance: $k(s, t) = \min\{s, t\} - st$



Example 2: Brownian Bridge

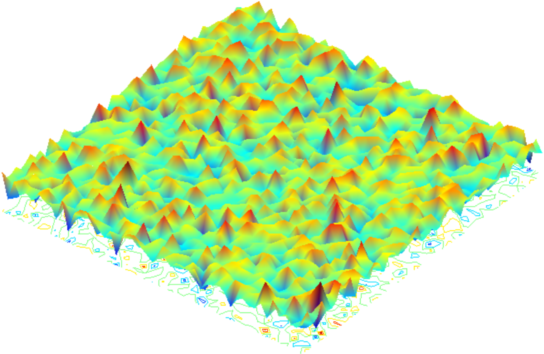
- A bridge is a stochastic process that is "clamped" at two points, i.e. each path goes (w.p. 1) through two specified points.
- Example: Brownian Bridge on $[0, 1]$ with $f(0) = f(1) = 0$.
- Covariance: $k(s, t) = \min\{s, t\} - st$



Example 3: Gauss Covariance

- Gauss covariance function:

$$k(x, y) = \exp\left(-\frac{1}{2\sigma}||x - y||_2^2\right).$$



Continuity and Differentiability of Sample Paths

- These three processes have continuous sample paths (w.p. 1).
- The process with the Gauss covariance has furthermore sample paths that are infinitely often differentiable (w.p. 1).
- Sample paths of Markov processes are very “rough” with a lot of fluctuations. The sample paths of Brownian motion are, for example, nowhere differentiable (w.p. 1).
- It is useful for modelling purposes to be able to specify the smoothnes of a process in terms of how often the sample paths are differentiable. The Matérn class of covariance functions allows to do that.

Continuity and Differentiability of Sample Paths

- These three processes have continuous sample paths (w.p. 1).
- The process with the Gauss covariance has furthermore sample paths that are infinitely often differentiable (w.p. 1).
- Sample paths of Markov processes are very “rough” with a lot of fluctuations. The sample paths of Brownian motion are, for example, nowhere differentiable (w.p. 1).
- It is useful for modelling purposes to be able to specify the smoothnes of a process in terms of how often the sample paths are differentiable. The Matérn class of covariance functions allows to do that.

Continuity and Differentiability of Sample Paths

- These three processes have continuous sample paths (w.p. 1).
- The process with the Gauss covariance has furthermore sample paths that are infinitely often differentiable (w.p. 1).
- Sample paths of Markov processes are very “rough” with a lot of fluctuations. The sample paths of Brownian motion are, for example, nowhere differentiable (w.p. 1).
- It is useful for modelling purposes to be able to specify the smoothnes of a process in terms of how often the sample paths are differentiable. The Matérn class of covariance functions allows to do that.

Continuity and Differentiability of Sample Paths

- These three processes have continuous sample paths (w.p. 1).
- The process with the Gauss covariance has furthermore sample paths that are infinitely often differentiable (w.p. 1).
- Sample paths of Markov processes are very “rough” with a lot of fluctuations. The sample paths of Brownian motion are, for example, nowhere differentiable (w.p. 1).
- It is useful for modelling purposes to be able to specify the smoothnes of a process in terms of how often the sample paths are differentiable. The Matérn class of covariance functions allows to do that.

GP Prediction - A bit of history ...

Definition (Kolmogorov-Wiener prediction prob. (1941))

Given a zero mean GP on the real line with covariance function k. What is the best prediction for the value of the process at time $\tau > 0$ given you observed the process on $(-\infty, 0]$.

- Leads to the so called Wiener filter.
- *The original motivation from Wiener was the targeting of air planes.*
- The prediction problem involving a continuum of observations is difficult and a deep theory is underlying it.
- Small changes of the setting can make things significantly more difficult. E.g. assume that you observe the process only on a finite past $(-T, 0]$. A completely different technique is needed to solve this problem.
- Still a topic with active research. The optimal filter can currently only be computed in special cases.

GP Prediction - A bit of history ...

UCL

Definition (Kolmogorov-Wiener prediction prob. (1941))

Given a zero mean GP on the real line with covariance function k. What is the best prediction for the value of the process at time $\tau > 0$ given you observed the process on $(-\infty, 0]$.

- Leads to the so called Wiener filter.
- *The original motivation from Wiener was the targeting of air planes.*
- The prediction problem involving a continuum of observations is difficult and a deep theory is underlying it.
- Small changes of the setting can make things significantly more difficult. E.g. assume that you observe the process only on a finite past $(-T, 0]$. A completely different technique is needed to solve this problem.
- Still a topic with active research. The optimal filter can currently only be computed in special cases.

GP Prediction - A bit of history ...

Definition (Kolmogorov-Wiener prediction prob. (1941))

Given a zero mean GP on the real line with covariance function k. What is the best prediction for the value of the process at time $\tau > 0$ given you observed the process on $(-\infty, 0]$.

- Leads to the so called Wiener filter.
- *The original motivation from Wiener was the targeting of air planes.*
- The prediction problem involving a continuum of observations is difficult and a deep theory is underlying it.
- Small changes of the setting can make things significantly more difficult. E.g. assume that you observe the process only on a finite past $(-T, 0]$. A completely different technique is needed to solve this problem.
- Still a topic with active research. The optimal filter can currently only be computed in special cases.

GP Prediction - A bit of history ...

Definition (Kolmogorov-Wiener prediction prob. (1941))

Given a zero mean GP on the real line with covariance function k. What is the best prediction for the value of the process at time $\tau > 0$ given you observed the process on $(-\infty, 0]$.

- Leads to the so called Wiener filter.
- *The original motivation from Wiener was the targeting of air planes.*
- The prediction problem involving a continuum of observations is difficult and a deep theory is underlying it.
- Small changes of the setting can make things significantly more difficult. E.g. assume that you observe the process only on a finite past $(-T, 0]$. A completely different technique is needed to solve this problem.
- Still a topic with active research. The optimal filter can currently only be computed in special cases.

GP Prediction - A bit of history ...

UCL

Definition (Kolmogorov-Wiener prediction prob. (1941))

Given a zero mean GP on the real line with covariance function k. What is the best prediction for the value of the process at time $\tau > 0$ given you observed the process on $(-\infty, 0]$.

- Leads to the so called Wiener filter.
- *The original motivation from Wiener was the targeting of air planes.*
- The prediction problem involving a continuum of observations is difficult and a deep theory is underlying it.
- Small changes of the setting can make things significantly more difficult. E.g. assume that you observe the process only on a finite past $(-T, 0]$. A completely different technique is needed to solve this problem.
- Still a topic with active research. The optimal filter can currently only be computed in special cases.

GP Prediction - A bit of history ...

Definition (Kolmogorov-Wiener prediction prob. (1941))

Given a zero mean GP on the real line with covariance function k. What is the best prediction for the value of the process at time $\tau > 0$ given you observed the process on $(-\infty, 0]$.

- Leads to the so called Wiener filter.
- *The original motivation from Wiener was the targeting of air planes.*
- The prediction problem involving a continuum of observations is difficult and a deep theory is underlying it.
- Small changes of the setting can make things significantly more difficult. E.g. assume that you observe the process only on a finite past $(-T, 0]$. A completely different technique is needed to solve this problem.
- Still a topic with active research. The optimal filter can currently only be computed in special cases.

UCL

- Another milestone: Kalman filter (1960)
- The Kalman filter approaches the problem differently:
  - The Wiener filter uses the covariance function to construct the optimal prediction.
  - The Kalman filter uses a state-space model.
  - It is easy to get the covariance from a state-space model.
  - But it is difficult to construct a suitable state-space model for a given covariance.
- The Kalman filter is an efficient approach to solve the prediction problem, but you need a state-space description.

GP Prediction - A bit of history ...

- Another milestone: Kalman filter (1960)
- The Kalman filter approaches the problem differently:
  - The Wiener filter uses the covariance function to construct the optimal prediction.
  - The Kalman filter uses a state-space model.
  - It is easy to get the covariance from a state-space model.
  - But it is difficult to construct a suitable state-space model for a given covariance.
- The Kalman filter is an efficient approach to solve the prediction problem, but you need a state-space description.

GP Prediction - A bit of history ...

- Another milestone: Kalman filter (1960)
- The Kalman filter approaches the problem differently:
  - The Wiener filter uses the covariance function to construct the optimal prediction.
  - The Kalman filter uses a state-space model.
  - It is easy to get the covariance from a state-space model.
  - But it is difficult to construct a suitable state-space model for a given covariance.
- The Kalman filter is an efficient approach to solve the prediction problem, but you need a state-space description.

GP Prediction - A bit of history ...

- Another milestone: Kalman filter (1960)
- The Kalman filter approaches the problem differently:
  - The Wiener filter uses the covariance function to construct the optimal prediction.
  - The Kalman filter uses a state-space model.
  - It is easy to get the covariance from a state-space model.
  - But it is difficult to construct a suitable state-space model for a given covariance.
- The Kalman filter is an efficient approach to solve the prediction problem, but you need a state-space description.

GP Prediction - A bit of history ...

- Another milestone: Kalman filter (1960)
- The Kalman filter approaches the problem differently:
  - The Wiener filter uses the covariance function to construct the optimal prediction.
  - The Kalman filter uses a state-space model.
  - It is easy to get the covariance from a state-space model.
  - But it is difficult to construct a suitable state-space model for a given covariance.
- The Kalman filter is an efficient approach to solve the prediction problem, but you need a state-space description.

GP Prediction - A bit of history ...

- Another milestone: Kalman filter (1960)
- The Kalman filter approaches the problem differently:
  - The Wiener filter uses the covariance function to construct the optimal prediction.
  - The Kalman filter uses a state-space model.
  - It is easy to get the covariance from a state-space model.
  - But it is difficult to construct a suitable state-space model for a given covariance.
- The Kalman filter is an efficient approach to solve the prediction problem, but you need a state-space description.

GP Prediction - A bit of history ...

- Another milestone: Kalman filter (1960)
- The Kalman filter approaches the problem differently:
  - The Wiener filter uses the covariance function to construct the optimal prediction.
  - The Kalman filter uses a state-space model.
  - It is easy to get the covariance from a state-space model.
  - But it is difficult to construct a suitable state-space model for a given covariance.
- The Kalman filter is an efficient approach to solve the prediction problem, but you need a state-space description.

GP Prediction - A bit of history ...

- During the “space age” a tremendous amount of money was spent on Kalman filter research.
- Quotes from Wiki ...:
  It was during a visit of Kalman to the NASA Ames Research Center that he saw the applicability of his ideas to the problem of trajectory estimation for the Apollo program, leading to its incorporation in the Apollo navigation computer.
- And:
  Kalman filters have been vital in the implementation of the navigation systems of U.S. Navy nuclear ballistic missile submarines; and in the guidance and navigation systems of cruise missiles such as the U.S. Navy’s Tomahawk missile; the U.S. Air Force’s Air Launched Cruise Missile; It is also used in the guidance and navigation systems of the NASA Space Shuttle and the attitude control and navigation systems of the International Space Station.

GP Prediction - A bit of history ...

- During the "space age" a tremendous amount of money was spent on Kalman filter research.
- Quotes from Wiki ...:
  It was during a visit of Kalman to the NASA Ames Research Center that he saw the applicability of his ideas to the problem of trajectory estimation for the Apollo program, leading to its incorporation in the Apollo navigation computer.
- And:
  Kalman filters have been vital in the implementation of the navigation systems of U.S. Navy nuclear ballistic missile submarines; and in the guidance and navigation systems of cruise missiles such as the U.S. Navy's Tomahawk missile; the U.S. Air Force's Air Launched Cruise Missile; It is also used in the guidance and navigation systems of the NASA Space Shuttle and the attitude control and navigation systems of the International Space Station.

GP Prediction - A bit of history ...

- During the “space age” a tremendous amount of money was spent on Kalman filter research.
- Quotes from Wiki ...:
  It was during a visit of Kalman to the NASA Ames Research Center that he saw the applicability of his ideas to the problem of trajectory estimation for the Apollo program, leading to its incorporation in the Apollo navigation computer.
- And:
  Kalman filters have been vital in the implementation of the navigation systems of U.S. Navy nuclear ballistic missile submarines; and in the guidance and navigation systems of cruise missiles such as the U.S. Navy’s Tomahawk missile; the U.S. Air Force’s Air Launched Cruise Missile; It is also used in the guidance and navigation systems of the NASA Space Shuttle and the attitude control and navigation systems of the International Space Station.

GP Prediction - A bit of history ...

- Two other fields where GP prediction has a long history are geostatistics (1973) and meterology (1956).
- In geostatistics it is known under the name kriging.
- In these fields GP prediction was naturally restricted to 2 and 3 dimensional input spaces.
- In the 90s people began to use GPs in machine learning.

GP Prediction - A bit of history ...

- Two other fields where GP prediction has a long history are geostatistics (1973) and meterology (1956).
- In geostatistics it is known under the name kriging.
- In these fields GP prediction was naturally restricted to 2 and 3 dimensional input spaces.
- In the 90s people began to use GPs in machine learning.

GP Prediction - A bit of history ...

- Two other fields where GP prediction has a long history are geostatistics (1973) and meterology (1956).
- In geostatistics it is known under the name kriging.
- In these fields GP prediction was naturally restricted to 2 and 3 dimensional input spaces.
- In the 90s people began to use GPs in machine learning.

GP Prediction - A bit of history ...

- Two other fields where GP prediction has a long history are geostatistics (1973) and meterology (1956).
- In geostatistics it is known under the name kriging.
- In these fields GP prediction was naturally restricted to 2 and 3 dimensional input spaces.
- In the 90s people began to use GPs in machine learning.

GP Regression

- Bayesian assumption - our function is drawn from a GP:

$$f(x) \sim \mathcal{GP}(\mu(x), k(x, y)).$$

- Remark: Distribution on a function space!
- Observation model:

$$y(x) = f(x) + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma_n)$ is observation noise.

- Posterior process for m observations $u_1, \dots, u_m$ (no continuum of observations):

$$f_{post}(z) \sim \mathcal{GP}(\mu_{post}, \boldsymbol{K}_{post})$$

$$\mu_{post}(z) = k(z, \boldsymbol{u})^\top (\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} y$$

$$\boldsymbol{K}_{post} = k(z, z) - k(z, \boldsymbol{u})(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} k(\boldsymbol{u}, z),$$

where $k(z, \boldsymbol{u}) = (k(z, u_1), \dots, k(z, u_m))$.

GP Regression

- Bayesian assumption - our function is drawn from a GP:

$$f(x) \sim \mathcal{GP}(\mu(x), k(x, y)).$$

- Remark: Distribution on a function space!
- Observation model:

$$y(x) = f(x) + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma_n)$ is observation noise.
- Posterior process for m observations $u_1, \dots, u_m$ (no continuum of observations):

$$f_{post}(z) \sim \mathcal{GP}(\mu_{post}, \boldsymbol{K}_{post})$$
$$\mu_{post}(z) = k(z, \boldsymbol{u})^\top (\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} y$$
$$\boldsymbol{K}_{post} = k(z, z) - k(z, \boldsymbol{u})(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} k(\boldsymbol{u}, z),$$

where $k(z, \boldsymbol{u}) = (k(z, u_1), \dots, k(z, u_m))$.

GP Regression

- Bayesian assumption - our function is drawn from a GP:

$$f(x) \sim \mathcal{GP}(\mu(x), k(x, y)).$$

- Remark: Distribution on a function space!
- Observation model:

$$y(x) = f(x) + \epsilon,$$

  where $\epsilon \sim \mathcal{N}(0, \sigma_n)$ is observation noise.
- Posterior process for m observations $u_1, \ldots, u_m$ (no continuum of observations):

$$f_{post}(z) \sim \mathcal{GP}(\mu_{post}, \boldsymbol{K}_{post})$$

$$\mu_{post}(z) = k(z, \boldsymbol{u})^\top (\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} y$$

$$\boldsymbol{K}_{post} = k(z, z) - k(z, \boldsymbol{u})(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} k(\boldsymbol{u}, z),$$

  where $k(z, \boldsymbol{u}) = (k(z, u_1), \ldots, k(z, u_m))$.

GP Regression

- Bayesian assumption - our function is drawn from a GP:

$$f(x) \sim \mathcal{GP}(\mu(x), k(x, y)).$$

- Remark: Distribution on a function space!
- Observation model:

$$y(x) = f(x) + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma_n)$ is observation noise.
- Posterior process for m observations $u_1, \ldots, u_m$ (no continuum of observations):

$$f_{post}(z) \sim \mathcal{GP}(\mu_{post}, \boldsymbol{K}_{post})$$
$$\mu_{post}(z) = k(z, \boldsymbol{u})^\top (\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} y$$
$$\boldsymbol{K}_{post} = k(z, z) - k(z, \boldsymbol{u})(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} k(\boldsymbol{u}, z),$$

where $k(z, \boldsymbol{u}) = (k(z, u_1), \ldots, k(z, u_m))$.

GP Regression - Relation to other Methods

- In the frequentiest world there exists a method called Ridge Regression. In the linear case the idea is to solve:

$$\min_{w} ||Aw - y||^2 + \lambda ||w||^2.$$

- There exists a kernel version and its solution is equivalent to the mean function of the GP.
- λ is in Ridge Regression a regularizer. In the GP setting this is the observation noise.
- Also very similar to Support Vector Regression.
- Difference: The Bayesian setting gives “error bars”, i.e. the variance estimate.
- However, these are no “true” error bars as they hold only under the Bayesian assumption (which is rarely fulfilled).
- The error bars does not depend on the concrete observations y, but only on the position of the observations and on the number of observations.

GP Regression - Relation to other Methods

- In the frequentiest world there exists a method called Ridge Regression. In the linear case the idea is to solve:

$$\min_w ||Aw - y||^2 + \lambda ||w||^2.$$

- There exists a kernel version and its solution is equivalent to the mean function of the GP.
- λ is in Ridge Regression a regularizer. In the GP setting this is the observation noise.
- Also very similar to Support Vector Regression.
- Difference: The Bayesian setting gives "error bars", i.e. the variance estimate.
- However, these are no "true" error bars as they hold only under the Bayesian assumption (which is rarely fulfilled).
- The error bars does not depend on the concrete observations y, but only on the position of the observations and on the number of observations.

GP Regression - Relation to other Methods

- In the frequentiest world there exists a method called Ridge Regression. In the linear case the idea is to solve:

$$\min_{w} ||Aw - y||^2 + \lambda ||w||^2.$$

- There exists a kernel version and its solution is equivalent to the mean function of the GP.
- λ is in Ridge Regression a regularizer. In the GP setting this is the observation noise.
- Also very similar to Support Vector Regression.
- Difference: The Bayesian setting gives “error bars”, i.e. the variance estimate.
- However, these are no “true” error bars as they hold only under the Bayesian assumption (which is rarely fulfilled).
- The error bars does not depend on the concrete observations y, but only on the position of the observations and on the number of observations.

GP Regression - Relation to other Methods

UCL

- In the frequentiest world there exists a method called Ridge Regression. In the linear case the idea is to solve:

$$\min_{w} ||Aw - y||^2 + \lambda ||w||^2.$$

- There exists a kernel version and its solution is equivalent to the mean function of the GP.
- λ is in Ridge Regression a regularizer. In the GP setting this is the observation noise.
- Also very similar to Support Vector Regression.
- Difference: The Bayesian setting gives "error bars", i.e. the variance estimate.
- However, these are no "true" error bars as they hold only under the Bayesian assumption (which is rarely fulfilled).
- The error bars does not depend on the concrete observations y, but only on the position of the observations and on the number of observations.

GP Regression - Relation to other Methods

- In the frequentiest world there exists a method called Ridge Regression. In the linear case the idea is to solve:

$$\min_{w} ||Aw - y||^2 + \lambda ||w||^2.$$

- There exists a kernel version and its solution is equivalent to the mean function of the GP.
- λ is in Ridge Regression a regularizer. In the GP setting this is the observation noise.
- Also very similar to Support Vector Regression.
- Difference: The Bayesian setting gives “error bars”, i.e. the variance estimate.
- However, these are no “true” error bars as they hold only under the Bayesian assumption (which is rarely fulfilled).
- The error bars does not depend on the concrete observations y, but only on the position of the observations and on the number of observations.

GP Regression - Relation to other Methods

- In the frequentiest world there exists a method called Ridge Regression. In the linear case the idea is to solve:

$$\min_w ||Aw - y||^2 + \lambda ||w||^2.$$

- There exists a kernel version and its solution is equivalent to the mean function of the GP.
- λ is in Ridge Regression a regularizer. In the GP setting this is the observation noise.
- Also very similar to Support Vector Regression.
- Difference: The Bayesian setting gives “error bars”, i.e. the variance estimate.
- However, these are no “true” error bars as they hold only under the Bayesian assumption (which is rarely fulfilled).
- The error bars does not depend on the concrete observations y, but only on the position of the observations and on the number of observations.

GP Regression - Relation to other Methods

- In the frequentiest world there exists a method called Ridge Regression. In the linear case the idea is to solve:

$$\min_{w} ||Aw - y||^2 + \lambda ||w||^2.$$

- There exists a kernel version and its solution is equivalent to the mean function of the GP.
- λ is in Ridge Regression a regularizer. In the GP setting this is the observation noise.
- Also very similar to Support Vector Regression.
- Difference: The Bayesian setting gives “error bars”, i.e. the variance estimate.
- However, these are no “true” error bars as they hold only under the Bayesian assumption (which is rarely fulfilled).
- The error bars does not depend on the concrete observations y, but only on the position of the observations and on the number of observations.

- GPs are also extensively studied in probability theory. Interesting are, for example, properties that hold for paths (w.p. 1).
- *Warm up:* Brownian motion.
- We stated already two important properties.
  - The sample paths are continuous (w.p. 1).
  - They are nowhere differentiable (w.p. 1).
- Bounds on the maximum of a Brownian motion:

$$\mathbf{P}[\sup_{u\in[0,t]} |f(u)| \geq b] \leq \sqrt{\frac{t}{2\pi}}\frac{4}{b}\exp\left(-\frac{b^2}{2t}\right).$$

- *Strong law of large numbers:*

$$\lim_{t\to\infty}\frac{f(t)}{t} = 0 \quad \text{(w.p. 1)}.$$

- How much does the paths oscilliate:

A sample path of a Brownian Motion is in no interval monotone (w.p. 1).

Pathwise Properties of GPs: Brownian Motion

- GPs are also extensively studied in probability theory. Interesting are, for example, properties that hold for paths (w.p. 1).
- *Warm up:* Brownian motion.
- We stated already two important properties.
  - The sample paths are continuous (w.p. 1).
  - They are nowhere differentiable (w.p. 1).
- Bounds on the maximum of a Brownian motion:

$$\mathbf{P}[\sup_{u \in [0,t]} |f(u)| \geq b] \leq \sqrt{\frac{t}{2\pi}} \frac{4}{b} \exp\left(-\frac{b^2}{2t}\right).$$

- *Strong law of large numbers:*

$$\lim_{t \to \infty} \frac{f(t)}{t} = 0 \quad \text{(w.p. 1)}.$$

- How much does the paths oscilliate:

  A sample path of a Brownian Motion is in no interval monotone (w.p. 1).

Pathwise Properties of GPs: Brownian Motion

- GPs are also extensively studied in probability theory. Interesting are, for example, properties that hold for paths (w.p. 1).
- *Warm up:* Brownian motion.
- We stated already two important properties.
  - The sample paths are continuous (w.p. 1).
  - They are nowhere differentiable (w.p. 1).
- Bounds on the maximum of a Brownian motion:

$$\mathbf{P}[\sup_{u\in[0,t]} |f(u)| \geq b] \leq \sqrt{\frac{t}{2\pi}}\frac{4}{b}\exp\left(-\frac{b^2}{2t}\right).$$

- *Strong law of large numbers:*

$$\lim_{t\to\infty} \frac{f(t)}{t} = 0 \quad \text{(w.p. 1)}.$$

- How much does the paths oscilliate:

A sample path of a Brownian Motion is in no interval monotone (w.p. 1).

- GPs are also extensively studied in probability theory. Interesting are, for example, properties that hold for paths (w.p. 1).
- *Warm up:* Brownian motion.
- We stated already two important properties.
  - The sample paths are continuous (w.p. 1).
  - They are nowhere differentiable (w.p. 1).
- Bounds on the maximum of a Brownian motion:

$$\mathbf{P}[\sup_{u\in[0,t]} |f(u)| \geq b] \leq \sqrt{\frac{t}{2\pi}}\frac{4}{b}\exp\left(-\frac{b^2}{2t}\right).$$

- *Strong law of large numbers:*

$$\lim_{t\to\infty}\frac{f(t)}{t} = 0 \quad \text{(w.p. 1)}.$$

- How much does the paths oscilliate:

A sample path of a Brownian Motion is in no interval monotone (w.p. 1).

Pathwise Properties of GPs: Brownian Motion

- GPs are also extensively studied in probability theory. Interesting are, for example, properties that hold for paths (w.p. 1).
- *Warm up:* Brownian motion.
- We stated already two important properties.
  - The sample paths are continuous (w.p. 1).
  - They are nowhere differentiable (w.p. 1).
- Bounds on the maximum of a Brownian motion:

$$\mathbf{P}[\sup_{u \in [0,t]} |f(u)| \geq b] \leq \sqrt{\frac{t}{2\pi}} \frac{4}{b} \exp\left(-\frac{b^2}{2t}\right).$$

- *Strong law of large numbers:*

$$\lim_{t \to \infty} \frac{f(t)}{t} = 0 \quad \text{(w.p. 1)}.$$

- How much does the paths oscilliate:

A sample path of a Brownian Motion is in no interval monotone (w.p. 1).

- GPs are also extensively studied in probability theory. Interesting are, for example, properties that hold for paths (w.p. 1).
- *Warm up:* Brownian motion.
- We stated already two important properties.
  - The sample paths are continuous (w.p. 1).
  - They are nowhere differentiable (w.p. 1).
- Bounds on the maximum of a Brownian motion:

$$\mathbf{P}[\sup_{u\in[0,t]} |f(u)| \geq b] \leq \sqrt{\frac{t}{2\pi}}\frac{4}{b}\exp\left(-\frac{b^2}{2t}\right).$$

- *Strong law of large numbers:*

$$\lim_{t\to\infty} \frac{f(t)}{t} = 0 \quad \text{(w.p. 1)}.$$

- How much does the paths oscilliate:

A sample path of a Brownian Motion is in no interval monotone (w.p. 1).

- GPs are also extensively studied in probability theory. Interesting are, for example, properties that hold for paths (w.p. 1).
- *Warm up:* Brownian motion.
- We stated already two important properties.
  - The sample paths are continuous (w.p. 1).
  - They are nowhere differentiable (w.p. 1).
- Bounds on the maximum of a Brownian motion:

$$\mathbf{P}[\sup_{u\in[0,t]} |f(u)| \geq b] \leq \sqrt{\frac{t}{2\pi}}\frac{4}{b}\exp\left(-\frac{b^2}{2t}\right).$$

- *Strong law of large numbers:*

$$\lim_{t\to\infty}\frac{f(t)}{t} = 0 \quad \text{(w.p. 1)}.$$

- How much does the paths oscilliate:

A sample path of a Brownian Motion is in no interval monotone (w.p. 1).

Pathwise Properties of GPs: General Case

- What about the sample paths of a GP with a given covariance k?
- It seems hopeless at the first look to get results for the general case.
- But deep results exist (!) that allow us to infer properties of sample paths of general GPs on a *compact input space*.
- Milestone: R.M. Dudley 1967 gave a criterion for sample path continuity and a way to bound the supremum of a GP on its input space.
- His criterion is *sufficient* for a GP to be continuous, but not *necessary*.
- Later Fernique and Talagrand derived sufficient and necessary criterion.
- In particular, Talagrand managed to get upper and lower bounds on the expected supremum of a GP on the same order (up to a constant).

Pathwise Properties of GPs: General Case

- What about the sample paths of a GP with a given covariance k?
- It seems hopeless at the first look to get results for the general case.
- But deep results exist (!) that allow us to infer properties of sample paths of general GPs on a *compact input space*.
- Milestone: R.M. Dudley 1967 gave a criterion for sample path continuity and a way to bound the supremum of a GP on its input space.
- His criterion is *sufficient* for a GP to be continuous, but not *necessary*.
- Later Fernique and Talagrand derived sufficient and necessary criterion.
- In particular, Talagrand managed to get upper and lower bounds on the expected supremum of a GP on the same order (up to a constant).

Pathwise Properties of GPs: General Case

- What about the sample paths of a GP with a given covariance k?
- It seems hopeless at the first look to get results for the general case.
- But deep results exist (!) that allow us to infer properties of sample paths of general GPs on a *compact input space*.
- Milestone: R.M. Dudley 1967 gave a criterion for sample path continuity and a way to bound the supremum of a GP on its input space.
- His criterion is *sufficient* for a GP to be continuous, but not *necessary*.
- Later Fernique and Talagrand derived sufficient and necessary criterion.
- In particular, Talagrand managed to get upper and lower bounds on the expected supremum of a GP on the same order (up to a constant).

Pathwise Properties of GPs: General Case

- What about the sample paths of a GP with a given covariance k?
- It seems hopeless at the first look to get results for the general case.
- But deep results exist (!) that allow us to infer properties of sample paths of general GPs on a *compact input space*.
- Milestone: R.M. Dudley 1967 gave a criterion for sample path continuity and a way to bound the supremum of a GP on its input space.
- His criterion is *sufficient* for a GP to be continuous, but not *necessary*.
- Later Fernique and Talagrand derived sufficient and necessary criterion.
- In particular, Talagrand managed to get upper and lower bounds on the expected supremum of a GP on the same order (up to a constant).

Pathwise Properties of GPs: General Case

- What about the sample paths of a GP with a given covariance k?
- It seems hopeless at the first look to get results for the general case.
- But deep results exist (!) that allow us to infer properties of sample paths of general GPs on a *compact input space*.
- Milestone: R.M. Dudley 1967 gave a criterion for sample path continuity and a way to bound the supremum of a GP on its input space.
- His criterion is *sufficient* for a GP to be continuous, but not *necessary*.
- Later Fernique and Talagrand derived sufficient and necessary criterion.
- In particular, Talagrand managed to get upper and lower bounds on the expected supremum of a GP on the same order (up to a constant).

Pathwise Properties of GPs: General Case

- What about the sample paths of a GP with a given covariance k?
- It seems hopeless at the first look to get results for the general case.
- But deep results exist (!) that allow us to infer properties of sample paths of general GPs on a *compact input space*.
- Milestone: R.M. Dudley 1967 gave a criterion for sample path continuity and a way to bound the supremum of a GP on its input space.
- His criterion is *sufficient* for a GP to be continuous, but not *necessary*.
- Later Fernique and Talagrand derived sufficient and necessary criterion.
- In particular, Talagrand managed to get upper and lower bounds on the expected supremum of a GP on the same order (up to a constant).

Pathwise Properties of GPs: General Case

- What about the sample paths of a GP with a given covariance k?
- It seems hopeless at the first look to get results for the general case.
- But deep results exist (!) that allow us to infer properties of sample paths of general GPs on a *compact input space*.
- Milestone: R.M. Dudley 1967 gave a criterion for sample path continuity and a way to bound the supremum of a GP on its input space.
- His criterion is *sufficient* for a GP to be continuous, but not *necessary*.
- Later Fernique and Talagrand derived sufficient and necessary criterion.
- In particular, Talagrand managed to get upper and lower bounds on the expected supremum of a GP on the same order (up to a constant).

Generic Chaining

- The original technique was quite demanding.
- Recently, Talagrand found a new technique called the *generic chaining*.
- How does it work:
- We assume in the following that the process is zero mean, i.e. $f \sim \mathcal{N}(0, k)$.
- One of the central ideas is to use a canonical metric for a GP. The canonical metric is:

$$d^2(x, y) = \mathbf{E}[(x - y)^2] = k(x, x) - 2k(x, y) + k(y, y).$$

- Remark: **the distance depends only on the covariance!**

Generic Chaining

- The original technique was quite demanding.
- Recently, Talagrand found a new technique called the *generic chaining*.
- How does it work:
- We assume in the following that the process is zero mean, i.e. $f \sim \mathcal{N}(0, k)$.
- One of the central ideas is to use a canonical metric for a GP. The canonical metric is:

$$d^2(x, y) = \mathbf{E}[(x - y)^2] = k(x, x) - 2k(x, y) + k(y, y).$$

- Remark: **the distance depends only on the covariance!**

Generic Chaining

- The original technique was quite demanding.
- Recently, Talagrand found a new technique called the *generic chaining*.
- How does it work:
- We assume in the following that the process is zero mean, i.e. $f \sim \mathcal{N}(0, k)$.
- One of the central ideas is to use a canonical metric for a GP. The canonical metric is:

$$d^2(x, y) = \mathbf{E}[(x - y)^2] = k(x, x) - 2k(x, y) + k(y, y).$$

- Remark: **the distance depends only on the covariance!**

Generic Chaining

- The original technique was quite demanding.
- Recently, Talagrand found a new technique called the *generic chaining*.
- How does it work:
- We assume in the following that the process is zero mean, i.e. $f \sim \mathcal{N}(0, k)$.
- One of the central ideas is to use a canonical metric for a GP. The canonical metric is:

$$d^2(x, y) = \mathbf{E}[(x - y)^2] = k(x, x) - 2k(x, y) + k(y, y).$$

- Remark: **the distance depends only on the covariance!**

Generic Chaining

- The original technique was quite demanding.
- Recently, Talagrand found a new technique called the *generic chaining*.
- How does it work:
- We assume in the following that the process is zero mean, i.e. $f \sim \mathcal{N}(0, k)$.
- One of the central ideas is to use a canonical metric for a GP. The canonical metric is:

$$d^2(x, y) = \mathbf{E}[(x - y)^2] = k(x, x) - 2k(x, y) + k(y, y).$$

- Remark: **the distance depends only on the covariance!**

Generic Chaining

- The idea is now to measure the size of the input space (let's call the space $\mathcal{X}$) with this canonical metric.
- The size is measured by partioning the space into N_n many parts, where

$$N_0 = 1 \quad \text{and} \quad N_n = 2^{2^n} \quad \text{if} \quad n > 0.$$

- We formalize this idea with the following definition:

Definition

Given a set $\mathcal{X}$ an admissible sequence is an increasing sequence $(\mathcal{A}_n)$ of partitions of $\mathcal{X}$ such that $\text{card}\mathcal{A}_n \leq N_n$.

Generic Chaining

- The idea is now to measure the size of the input space (let's call the space $\mathcal{X}$) with this canonical metric.
- The size is measured by partioning the space into N_n many parts, where

$$N_0 = 1 \quad \text{and} \quad N_n = 2^{2^n} \quad \text{if} \quad n > 0.$$

- We formalize this idea with the following definition:

Definition

Given a set $\mathcal{X}$ an admissible sequence is an increasing sequence $(\mathcal{A}_n)$ of partitions of $\mathcal{X}$ such that $\text{card}\mathcal{A}_n \leq N_n$.

- The idea is now to measure the size of the input space (let's call the space $\mathcal{X}$) with this canonical metric.
- The size is measured by partioning the space into N_n many parts, where

$$N_0 = 1 \quad \text{and} \quad N_n = 2^{2^n} \quad \text{if} \quad n > 0.$$

- We formalize this idea with the following definition:

Definition

Given a set $\mathcal{X}$ an admissible sequence is an increasing sequence $(\mathcal{A}_n)$ of partitions of $\mathcal{X}$ such that $\text{card}\mathcal{A}_n \leq N_n$.

UCL

- The idea is now to measure the size of the input space (let's call the space $\mathcal{X}$) with this canonical metric.
- The size is measured by partioning the space into N_n many parts, where

$$N_0 = 1 \quad \text{and} \quad N_n = 2^{2^n} \quad \text{if} \quad n > 0.$$

- We formalize this idea with the following definition:

Definition

Given a set $\mathcal{X}$ an admissible sequence is an increasing sequence $(\mathcal{A}_n)$ of partitions of $\mathcal{X}$ such that $\text{card}\mathcal{A}_n \leq N_n$.

Generic Chaining Bound

Theorem (The generic chaining bound.)

For a zero mean Gaussian process $f(x)$ we have for each admissible sequence that

$$\mathbf{E} \sup_{x \in \mathcal{X}} f(x) \leq 14 \sup_{x \in \mathcal{X}} \sum_{n \geq 0} 2^{n/2} \Delta(A_n(x)).$$

- Here, $A_n(x)$ is the set in the partition $\mathcal{A}_n$ in which x lies.
- $\Delta(A) = \sup_{x,y \in A} d(x,y)$ is the diameter of the set A measured with the canonical metric d.
- The interesting property is here that only analytic objects are involved - no stochastic elements are present!
- Furthermore: If $\mathbf{E} \sup_{x \in \mathcal{X}} f(x) < \infty$ then the GP has continuous sample paths (w.p. 1)!

Generic Chaining Bound

Theorem (The generic chaining bound.)

For a zero mean Gaussian process $f(x)$ we have for each admissible sequence that

$$\mathbf{E} \sup_{x \in \mathcal{X}} f(x) \leq 14 \sup_{x \in \mathcal{X}} \sum_{n \geq 0} 2^{n/2} \Delta(A_n(x)).$$

- Here, $A_n(x)$ is the set in the partition $\mathcal{A}_n$ in which x lies.
- $\Delta(A) = \sup_{x,y \in A} d(x,y)$ is the diameter of the set A measured with the canonical metric d.
- The interesting property is here that only analytic objects are involved - no stochastic elements are present!
- Furthermore: If $\mathbf{E} \sup_{x \in \mathcal{X}} f(x) < \infty$ then the GP has continuous sample paths (w.p. 1)!

Generic Chaining Bound

Theorem (The generic chaining bound.)

For a zero mean Gaussian process $f(x)$ we have for each admissible sequence that

$$\mathbf{E} \sup_{x \in \mathcal{X}} f(x) \leq 14 \sup_{x \in \mathcal{X}} \sum_{n \geq 0} 2^{n/2} \Delta(A_n(x)).$$

- Here, $A_n(x)$ is the set in the partition $\mathcal{A}_n$ in which x lies.
- $\Delta(A) = \sup_{x,y \in A} d(x, y)$ is the diameter of the set A measured with the canonical metric d.
- The interesting property is here that only analytic objects are involved - no stochastic elements are present!
- Furthermore: If $\mathbf{E} \sup_{x \in \mathcal{X}} f(x) < \infty$ then the GP has continuous sample paths (w.p. 1)!

Generic Chaining Bound

Theorem (The generic chaining bound.)

For a zero mean Gaussian process $f(x)$ we have for each admissible sequence that

$$\mathbf{E} \sup_{x \in \mathcal{X}} f(x) \leq 14 \sup_{x \in \mathcal{X}} \sum_{n \geq 0} 2^{n/2} \Delta(A_n(x)).$$

- Here, $A_n(x)$ is the set in the partition $\mathcal{A}_n$ in which x lies.
- $\Delta(A) = \sup_{x,y \in A} d(x, y)$ is the diameter of the set A measured with the canonical metric d.
- The interesting property is here that only analytic objects are involved - no stochastic elements are present!
- Furthermore: If $\mathbf{E} \sup_{x \in \mathcal{X}} f(x) < \infty$ then the GP has continuous sample paths (w.p. 1)!

Generic Chaining Bound

Theorem (The generic chaining bound.)

For a zero mean Gaussian process $f(x)$ we have for each admissible sequence that

$$\mathbf{E} \sup_{x \in \mathcal{X}} f(x) \leq 14 \sup_{x \in \mathcal{X}} \sum_{n \geq 0} 2^{n/2} \Delta(A_n(x)).$$

- Here, $A_n(x)$ is the set in the partition $\mathcal{A}_n$ in which x lies.
- $\Delta(A) = \sup_{x,y \in A} d(x,y)$ is the diameter of the set A measured with the canonical metric d.
- The interesting property is here that only analytic objects are involved - no stochastic elements are present!
- Furthermore: If $\mathbf{E} \sup_{x \in \mathcal{X}} f(x) < \infty$ then the GP has continuous sample paths (w.p. 1)!

Exercise (need a volunteer): Prove sample path continuity of the Brownian motion and derive a bound on its maximum!

Upper and Lower Bounds

- A corresponding lower bound exists. To state this we define:

Definition

Given an input space $\mathcal{X}$ and the canonical metric d then

$$\gamma_2(X, d) = \inf \sup_{x \in \mathcal{X}} \sum_{n \geq 0} 2^{n/2} \Delta(A_n(x)).$$

- Difference: the infimum is taken over all admissible sequences.
- $\gamma_2(\mathcal{X}, d)$ allows us to upper and lower bound the expected supremum:

$$\frac{1}{L}\gamma_2(\mathcal{X}, d) \leq \mathbf{E} \sup_{x \in \mathcal{X}} f(x) \leq L\gamma_2(\mathcal{X}, d).$$

Upper and Lower Bounds

- A corresponding lower bound exists. To state this we define:

Definition

Given an input space $\mathcal{X}$ and the canonical metric d then

$$\gamma_2(X, d) = \inf \sup_{x \in \mathcal{X}} \sum_{n \geq 0} 2^{n/2} \Delta(A_n(x)).$$

- Difference: the infimum is taken over all admissible sequences.
- $\gamma_2(\mathcal{X}, d)$ allows us to upper and lower bound the expected supremum:

$$\frac{1}{L}\gamma_2(\mathcal{X}, d) \leq \mathbf{E} \sup_{x \in \mathcal{X}} f(x) \leq L\gamma_2(\mathcal{X}, d).$$

Upper and Lower Bounds

- A corresponding lower bound exists. To state this we define:

Definition

Given an input space $\mathcal{X}$ and the canonical metric d then

$$\gamma_2(X, d) = \inf \sup_{x \in \mathcal{X}} \sum_{n \geq 0} 2^{n/2} \Delta(A_n(x)).$$

- Difference: the infimum is taken over all admissible sequences.
- $\gamma_2(\mathcal{X}, d)$ allows us to upper and lower bound the expected supremum:

$$\frac{1}{L}\gamma_2(\mathcal{X}, d) \leq \mathbf{E} \sup_{x \in \mathcal{X}} f(x) \leq L\gamma_2(\mathcal{X}, d).$$

Upper and Lower Bounds

- A corresponding lower bound exists. To state this we define:

Definition

Given an input space $\mathcal{X}$ and the canonical metric d then

$$\gamma_2(X, d) = \inf \sup_{x \in \mathcal{X}} \sum_{n \geq 0} 2^{n/2} \Delta(A_n(x)).$$

- Difference: the infimum is taken over all admissible sequences.
- $\gamma_2(\mathcal{X}, d)$ allows us to upper and lower bound the expected supremum:

$$\frac{1}{L}\gamma_2(\mathcal{X}, d) \leq \mathbf{E} \sup_{x \in \mathcal{X}} f(x) \leq L\gamma_2(\mathcal{X}, d).$$

Borell inequality

- There exist two important basic theorems for GPs:
  1. The Borell inequality.
  2. Slepian's inequality.
- Borell links the probability of a deviation to the expected supremum's bound:

Theorem (Borell inequality)

Let $f(x)$ be a centerd GP with sample paths being bounded w.p. 1. Let $||r|| = \sup_{x \in \mathcal{X}} r(x)$. Then

$$\mathbf{P}[|\,||r|| - \mathbf{E}||r||\,| > \lambda] \leq 2 \exp\left(-\frac{1}{2}\frac{\lambda^2}{\sigma_{\mathcal{X}}^2}\right).$$

Borell inequality

- There exist two important basic theorems for GPs:
  1. The Borell inequality.
  2. Slepian's inequality.
- Borell links the probability of a deviation to the expected supremum's bound:

Theorem (Borell inequality)

Let $f(x)$ be a centerd GP with sample paths being bounded w.p. 1. Let $||r|| = \sup_{x \in \mathcal{X}} r(x)$. Then

$$\mathbf{P}[|\,||r|| - \mathbf{E}||r||\,| > \lambda] \leq 2 \exp\left(-\frac{1}{2}\frac{\lambda^2}{\sigma_{\mathcal{X}}^2}\right).$$

Borell inequality

- There exist two important basic theorems for GPs:
  1. The Borell inequality.
  2. Slepian's inequality.
- Borell links the probability of a deviation to the expected supremum's bound:

Theorem (Borell inequality)

Let $f(x)$ be a centerd GP with sample paths being bounded w.p. 1. Let $||r|| = \sup_{x \in \mathcal{X}} r(x)$. Then

$$\mathbf{P}[|\,||r|| - \mathbf{E}||r||\,| > \lambda] \leq 2 \exp\left(-\frac{1}{2}\frac{\lambda^2}{\sigma_{\mathcal{X}}^2}\right).$$

Borell inequality

- There exist two important basic theorems for GPs:
  1. The Borell inequality.
  2. Slepian's inequality.
- Borell links the probability of a deviation to the expected supremum's bound:

Theorem (Borell inequality)

Let $f(x)$ be a centerd GP with sample paths being bounded w.p. 1. Let $||r|| = \sup_{x \in \mathcal{X}} r(x)$. Then

$$\mathbf{P}[|\,||r|| - \mathbf{E}||r||\,| > \lambda] \leq 2 \exp\left(-\frac{1}{2}\frac{\lambda^2}{\sigma_{\mathcal{X}}^2}\right).$$

Borell inequality

- There exist two important basic theorems for GPs:
  1. The Borell inequality.
  2. Slepian's inequality.
- Borell links the probability of a deviation to the expected supremum's bound:

Theorem (Borell inequality)

Let $f(x)$ be a centerd GP with sample paths being bounded w.p. 1. Let $||r|| = \sup_{x \in \mathcal{X}} r(x)$. Then

$$\mathbf{P}[|\,||r|| - \mathbf{E}||r||\,| > \lambda] \leq 2 \exp\left(-\frac{1}{2}\frac{\lambda^2}{\sigma_{\mathcal{X}}^2}\right).$$

Slepian's inequality

- Slepian's inequality is very intuitiv. It links the suprema distribution of two related GPs:

Theorem (Slepian's inequality)

Let $f(x)$ and $g(x)$ are centerd GPs with sample paths being bounded w.p. 1,

$$\mathbf{E}f(x)^2 = \mathbf{E}g(x)^2$$

and

$$\mathbf{E}(f(x) - f(y))^2 \leq \mathbf{E}(g(x) - g(y))^2$$

then for all λ:

$$\mathbf{P}[\sup_x f(x) > \lambda] \leq \mathbf{P}[\sup_x g(x) > \lambda].$$

Slepian's inequality

- Slepian's inequality is very intuitiv. It links the suprema distribution of two related GPs:

Theorem (Slepian's inequality)

Let $f(x)$ and $g(x)$ are centerd GPs with sample paths being bounded w.p. 1,

$$\mathbf{E}f(x)^2 = \mathbf{E}g(x)^2$$

and

$$\mathbf{E}(f(x) - f(y))^2 \leq \mathbf{E}(g(x) - g(y))^2$$

then for all λ:

$$\mathbf{P}[\sup_x f(x) > \lambda] \leq \mathbf{P}[\sup_x g(x) > \lambda].$$

- One application of the theory is to control the probability of rare events, like what is the probability that a river crosses a certain level.
- Rare events are also important for statistics, e.g. to bound the generalization error.
- Another application is global optimization and Bandit problems.
- Task: Find the optimum of a cost function where the cost function is drawn from a GP:

$$f(x) \sim \mathcal{GP}(0, k).$$

- Idea: Try a number of points and control the probability that the posterior process achieves a supremum greater than b.

- One application of the theory is to control the probability of rare events, like what is the probability that a river crosses a certain level.
- Rare events are also important for statistics, e.g. to bound the generalization error.
- Another application is global optimization and Bandit problems.
- Task: Find the optimum of a cost function where the cost function is drawn from a GP:

$$f(x) \sim \mathcal{GP}(0, k).$$

- Idea: Try a number of points and control the probability that the posterior process achieves a supremum greater than b.

- One application of the theory is to control the probability of rare events, like what is the probability that a river crosses a certain level.
- Rare events are also important for statistics, e.g. to bound the generalization error.
- Another application is global optimization and Bandit problems.
- Task: Find the optimum of a cost function where the cost function is drawn from a GP:

$$f(x) \sim \mathcal{GP}(0, k).$$

- Idea: Try a number of points and control the probability that the posterior process achieves a supremum greater than b.

- One application of the theory is to control the probability of rare events, like what is the probability that a river crosses a certain level.
- Rare events are also important for statistics, e.g. to bound the generalization error.
- Another application is global optimization and Bandit problems.
- Task: Find the optimum of a cost function where the cost function is drawn from a GP:

$$f(x) \sim \mathcal{GP}(0, k).$$

- Idea: Try a number of points and control the probability that the posterior process achieves a supremum greater than b.

- One application of the theory is to control the probability of rare events, like what is the probability that a river crosses a certain level.
- Rare events are also important for statistics, e.g. to bound the generalization error.
- Another application is global optimization and Bandit problems.
- Task: Find the optimum of a cost function where the cost function is drawn from a GP:

$$f(x) \sim \mathcal{GP}(0, k).$$

- Idea: Try a number of points and control the probability that the posterior process achieves a supremum greater than b.